Genomic Data and Privacy: Background and Relevant Law

Updated May 11, 2015

Congressional Research Service
https://crsreports.congress.gov
R44026

CRS REPORT
Prepared for Members and
Committees of Congress

Summary

Advances in genomics technology and information technology infrastructure, together with policies regarding the sharing of research data, support new approaches to genomic research but also raise new issues with respect to privacy. The development of new genomic sequencing technologies has allowed for the generation of big data, and recent changes in information technology infrastructure have facilitated big data storage and analytics. These developments are expected to support significant changes in health research and, eventually, in health care delivery.

Genetic and genomic research—and other "omics" research—have generated large amounts of genetic data. If these "large-scale genomic data" are generated as a part of research funded by the National Institutes of Health (NIH), then they are subject to specific data sharing policies and are often held in publicly available databases. Among other things, advances in sequencing technology have enabled this research, making large amounts of data available at a rate that has generally outpaced the ability to both store and analyze that data.

NIH has established a comprehensive policy for the sharing of genomic data that "applies to all NIH-funded research that generates large-scale human or non-human genomic data as well as the use of these data for subsequent research." This policy requires investigators to outline their data sharing plans as part of their funding applications; if investigators fail to submit the required data, NIH may withhold funding. Investigators are required to de-identify the data prior to submitting it to NIH-designated data repositories, according to the requirements of both the HHS Common Rule and the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule.

Some recent studies have begun to suggest that different types of molecular data may be more likely to cause privacy issues than had been previously understood, and specifically, that de-identified large-scale genomic sequence data may in fact be able to be reidentified. In a recent study, researchers were able to reidentify research participants using the publicly available de-identified personal genome data and other publicly available metadata.

This demonstration of reidentified individuals in a research study using de-identified genome data raises the question of whether—and if so, how—relevant current law should be modified in response to this new capability. Relevant law governs informed consent, access to research data, and the use of this data, and includes (1) the Health Insurance Portability and Accountability Act (HIPAA) Privacy and Security Rules; (2) the HHS Regulations for the Protection of Human Research Subjects, or the Common Rule; and (3) the Genetic Information Nondiscrimination Act of 2008 (GINA). In addition, the Freedom of Information Act (FOIA) is relevant, not in the sense that it protects information from a potential privacy breach, but in that it allows public access to much of the information held by the federal government. This report discusses these considerations in the context of each of the relevant laws and regulations.

Contents

Contacts

Introduction

Advances in genomics technology and information technology infrastructure, together with policies regarding the sharing of research data, support expanded genomic research efforts but also raise new issues with respect to privacy, and specifically the effort to balance "the potential of scientific progress with privacy and respect for persons."[1] The development of new genomic sequencing technologies has allowed for the generation of big data, and recent changes in information technology infrastructure—including, for example, cloud data storage—have facilitated big data storage and analytics. These developments are expected to support significant changes in health research and, eventually, in health care delivery.[2] Specifically, researchers hope to leverage big data by combining genetic, environmental, clinical, behavioral, and other data to facilitate precision medicine. Precision medicine is the idea of providing health care to individuals based on specific patient and disease characteristics, and is a priority of the National Institutes of Health (NIH).[3]

What Is DNA Sequencing?

DNA sequencing determines the order of the building blocks of DNA—called nucleotides, and abbreviated A,T,C and G—in an individual's genetic code.

Leading up to and during the Human Genome Project (HGP), new approaches were developed that allowed researchers to sequence large whole genomes, including a technology called "shotgun sequencing." In shotgun sequencing, DNA is broken up into random pieces, the pieces are sequenced, and then they are pieced back together using either their overlapping regions or a reference sequence.

Next generation sequencing (NGS) technologies build on these technologies by parallelizing sequencing and allowing for the simultaneous sequencing of thousands or millions of small pieces of DNA. NGS has reduced the time it takes to sequence large quantities of DNA and has significantly lowered the associated cost.

Sources: https://www.genome.gov/glossary/index.cfm?id=51; http://www.nature.com/subjects/next-generation-sequencing.

Genetic and genomic research have generated large amounts of genetic data. If these "large-scale genomic data" are generated as a part of research funded by the National Institutes of Health (NIH), then they are subject to specific data sharing policies and are often held in publicly available databases. "Large-scale genomic data," as defined for the purposes of NIH's data sharing policy, include genome-wide association studies (GWAS), [4] as well as genome sequence, gene expression, and other data.[5]

---

[1] Gutmann, A. and J. W. Wagner, "Found Your DNA on the Web: Reconciling Privacy and Progress," Hastings Center Report, May-June 2013, pp. 15-18.

[2] Roski J., G. W. Bo-Linn, T. A. Andrews, "Creating Value In Health Care Through Big Data: Opportunities And Policy Implications," *Health Affairs*, vol. 33, no. 7, pp. 1115-1122, 2014.

[3] The White House, Office of the Press Secretary, "Fact Sheet: President Obama's Precision medicine Initiative," January 30, 2015, http://www.whitehouse.gov/the-press-office/2015/01/30/fact-sheet-president-obama-s-precision-medicine-initiative.

[4] National Institutes of Health, "National Institutes of Health Genomic Data Sharing Policy," http://gds.nih.gov/PDF/NIH_GDS_Policy.pdf. GWAS are used to identify genetic changes that have only a modest effect on disease or other phenotypes. GWAS are formally defined as "a study in which the density of genetic markers and the extent of linkage disequilibrium should be sufficient to capture ... a large proportion of the common variation in the genome of the population under study, and the number of samples ... should provide sufficient power to detect variants of modest effect."

[5] National Human Genome Research Institute (NHGRI), "Talking Glossary of Genetic Terms," https://www.genome.gov/Glossary/index.cfm. Gene expression data is data about "the process by which the

Among other things, advances in sequencing technology have enabled this research, and have made available large amounts of data at a rate that has generally outpaced the ability to both store and analyze that data (see **"What is DNA Sequencing?"** text box, above). Sequencing output has been increasing at approximately a fivefold rate per year in recent years,[6] and the sequence data from a single individual's genome uses about 100 gigabytes of storage space.[7] As a result, "[g]enomic databases increasingly surpass the storage abilities of individual researchers—and even of large institutions—and consequently are stored increasingly frequently in the 'cloud.'"[8,9] For example, the National Cancer Institute (NCI) at NIH has launched and funded its Cancer Genomics Cloud Pilots, an initiative to develop up to three public cancer genomics cloud pilots, where large data repositories will be colocated with computing resources.[10] Storage of sequence data in "clouds" facilitates faster, more widespread, and increased access to this information.

Genomic Data Sharing

Advances in genomics research—for example, studying the genetic underpinnings of common diseases such as diabetes—have been facilitated by the data-banking of large quantities of data that are in turn available as a result of policies that encourage or require data sharing. NIH has established a comprehensive policy for the sharing of genomic data that "applies to all NIH-funded research that generates large-scale human or non-human genomic data as well as the use of these data for subsequent research."[11] This policy requires investigators to outline their data sharing plans as part of their funding applications; if investigators fail to submit the required data, NIH may withhold funding.

Investigators are required to de-identify the data prior to submitting it to NIH-designated data repositories. Data should be de-identified—stripped of identifiers such as an individual's name—according to the requirements of both the (1) HHS Common Rule[12] and (2) the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule.[13] In addition, NIH requires the

information encoded in a gene is used to direct the assembly of a protein molecule."

[6] M. C. Schatz, B. Langmead, and S. L. Salzberg, "Cloud Computing and the DNA Data Race," *Nature Biotechnology*, vol. 28, no. 7, July 2010, pp. 691-693.

[7] K. A. Philips, J. R. Trosman, R. K. Kelley, et al., "Genomic Sequencing: Assessing The Health Care System, Policy, And Big-Data Implications," *Health Affairs*, vol. 33, no. 7, July 2014, pp. 1246-1253.

[8] Gutmann, A. and J. W. Wagner, "Found Your DNA on the Web: Reconciling Privacy and Progress," Hastings Center Report, May-June 2013, pp. 15-18. The National Institute of Standards and Technology (NIST) defines cloud computing as "a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction." See http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf.

[9] The National Institute of Standards and Technology (NIST) defines cloud computing as "... a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction." See http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf.

[10] NCI, National Cancer Informatics Program, "NCI Cancer Genomics Cloud Pilots," https://cbiit.nci.nih.gov/ncip/nci-cancer-genomics-cloud-pilots.

[11] National Institutes of Health, "National Institutes of Health Genomic Data Sharing Policy," http://gds.nih.gov/PDF/NIH_GDS_Policy.pdf, p. 1.

[12] 45 C.F.R. Part 46, Subpart A.

[13] 45 C.F.R. Part 164, Subpart E.

assignment of random and unique codes to the data, with the key linking the codes with individual identifiers to be held by the submitting institution.

The data are submitted to relevant NIH-designated databases (e.g., NIH database of Genotypes and Phenotypes, or dbGaP[14]) by the investigator, and the institution's Institutional Review Board (IRB) determines whether the data may be held in unrestricted-access repositories or should be available only through controlled-access. This determination is made based on the informed consent under which the research was conducted and specifically language in the informed consent about the use of the data for future research purposes, as well as for broad data sharing purposes. Investigators wishing to use controlled-access data for the purposes of secondary research must first get NIH approval to use the data for their specific research project. This requirement is in contrast to unrestricted-access data, which are publicly available to anyone, without requiring prior approval for use in secondary research.

Some databases provide differing levels of access to data, depending on the type of data involved. For example, the dbGaP has both open- and controlled-access levels. This flexibility allows for the "broad release of non-sensitive data, while providing oversight and investigator accountability for sensitive data sets involving personal health information."[15]

De-identified Genomic Sequence Data and Privacy Considerations

Collecting the large quantities of data needed to answer questions about the genetic underpinning of common diseases, and to support precision medicine, requires individuals who are willing to participate in research studies. "The willingness of individuals and communities to assume some risk to participate in biomedical research depends on the scientific community's ability to maintain the public's trust."[16] This trust is developed in many ways, including honest and complete disclosure of risks upfront (through informed consent) and safeguards to protect privacy. With respect to research studies, privacy may be considered in terms of three components, or decision points: (1) the individual's decision to disclose personal data, (2) decisions about controls on access to the data, and (3) decisions about appropriate uses (and what constitutes "misuse") of the data.[17] Relevant law and regulation—including the HIPAA Privacy Rule, the Genetic Information Nondiscrimination Act (GINA), and the Common Rule—govern aspects of these components, including the informed consent process for human research subjects, prohibited uses of the data, and access to the data (and in what form).

To date, the privacy of data has been largely considered in the context of identifiability; that is, whether data or information may be readily linked with an individual. NIH defines de-identified data as that data where information that could be used to associate the data with an individual has been removed.[18] Where data is considered to be "de-identified," relevant laws and regulations generally treat it as not posing the potential for a breach of individual privacy. This issue is illustrated by large-scale genomic sequence data, which are generated in large quantities, often

[14] For a listing of NIH Data Repositories, NIH-Funded Databases, and NIH Database Collaborations, see http://gds.nih.gov/02dr2.html.

[15] NIH, "dbGaP Overview," http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/about.html.

[16] L. L. Rodriguez., L. D. Brooks, J. H. Greenberg, and E. D. Green, "The Complexities of Genomic Identifiability," Science, vol. 339, January 2013, pp. 275-276.

[17] Gutmann, A. and J. W. Wagner, "Found Your DNA on the Web: Reconciling Privacy and Progress," Hastings Center Report, May-June 2013, pp. 15-18.

[18] National Institutes of Health, "National Institutes of Health Genomic Data Sharing Policy," http://gds.nih.gov/PDF/NIH_GDS_Policy.pdf, p. 4.

entered into the public domain, and generally agreed to be de-identified in the absence of other data sources. In other words, such data were not able to be linked back to a specific individual, or to be "reidentified."

Reidentification of Individuals Using De-identified Genomic Sequence Data

Experts have noted that "[r]ecent work reveals the need to re-examine the current paradigms for managing the potential identifiability of genomic data."[19] Recent studies suggest that different types of genomic data may be more likely to raise privacy issues than had been previously understood.[20] One study, in particular, has challenged the traditional paradigm that de-identified personal genome data could not be reidentified. In January 2013, a study by Gymrek et al. was published wherein researchers were able to reidentify nearly 50 participants in the International HapMap Project.[21] Researchers were able to reidentify these individuals using their publicly available de-identified personal genome data and other publicly available data.[22]

Essentially, to reidentify de-identified personal genome data, a researcher needs to be able to match the de-identified data to a second source of genetic data that is, in turn, linked to some (or multiple) pieces of identifying data. The January 2013 study by Gymrek et al. accomplished this by using a public genealogy database containing genetic information linked to surnames as its second source of genetic data.[23] The study matched the de-identified personal genome data with the genetic information contained in the genealogy database, allowing the researchers to link the de-identified genome data with a surname and other identifying pieces of data (e.g., geographical location). The researchers then used public databases to combine surname, year of birth, and state of residence to identify specific individuals.

As noted, the 2013 Gymrek study relied on genealogy data. For this reason, the study's findings cannot be assumed to be broadly generalizable to the whole population;[24] that is, the risk of reidentification using this particular second source of genetic data may not be uniform throughout the general population. In addition, this study relied on the 1000 Genomes database, which is open access; conversely, most NIH data repositories maintain some or all data in a controlled-access manner, which makes accessing data more time-consuming and makes it more difficult to access multiple databases at one time.

However, genetic genealogy and other genomic databases are growing, and "the more genomic data collected, and the more refined the connections between genetic variations, disease states, and other personal characteristics, the easier it becomes to reidentify an individual and discover private information."[25] Given the technical feasibility of reidentification demonstrated by the 2013 Gymrek study, along with the increasing amount of genetic data and analytics available, it is

[19] L. L. Rodriguez, L. D. Brooks, J. H. Greenberg, E. D. Green, "The Complexities of Genomic Identifiability," *Science*, vol.339, pp. 275-276, January 18, 2013.

[20] L. L. Rodriguez, L. D. Brooks, J. H. Greenberg, E. D. Green, "The Complexities of Genomic Identifiability," *Science*, vol.339, pp. 275-276, January 18, 2013.

[21] NIH, "International HapMap Project," http://hapmap.ncbi.nlm.nih.gov/citinghapmap.html.en.

[22] M. Gymrek, A. L. McGuire, D. Golan, E. Halperin, and Y. Erlich, "Identifying Personal Genomes by Surname Inference," *Science*, vol. 339, pp. 321-324, January 18, 2013.

[23] An example of a public searchable genetic genealogy database is YSearch, http://www.ysearch.org/.

[24] L. L. Rodriguez, L. D. Brooks, J. H. Greenberg, E. D. Green, "The Complexities of Genomic Identifiability," *Science*, vol.339, pp. 275-276, January 18, 2013.

[25] Gutmann, A. and J. W. Wagner, "Found Your DNA on the Web: Reconciling Privacy and Progress," Hastings Center Report, May-June 2013, pp. 15-16.

reasonable to expect that, over time, the risk of reidentification will expand to encompass more of the general population.

Genomic Data Sharing and Current Law

The NIH and others see value in genomic research and prioritize funding it and widely sharing the resulting data. However, the generation, handling, and release of these data require privacy and security protections. The recent reidentification of research participants demonstrates that privacy and security concerns are not merely theoretical. Given this, policymakers may decide to monitor both NIH's evolving genomic data sharing policies and relevant federal law. Relevant law includes (1) the Health Insurance Portability and Accountability Act (HIPAA) Privacy and Security Rules; (2) the HHS Regulations for the Protection of Human Research Subjects, or the Common Rule; (3) GINA; and (4) the Freedom of Information Act (FOIA). FOIA is relevant, not in the sense that it protects information from a potential privacy breach, but in that it allows public access to much of the information held by the federal government. The remaining sections of this report provide an overview of each of these relevant laws and regulations.

Genomic Data Privacy and Security[26]

NIH's Genomic Data Sharing (GDS) Policy addresses both the submission of genomic data by NIH-funded researchers to an NIH data repository and the subsequent access and use of that data by other investigators.[27] Generally, NIH-funded researchers conducting human genomic research must adhere to the Common Rule[28] and have their studies approved by an Institutional Review Board (IRB). For research that falls under the scope of the GDS Policy, the IRB must review the informed consent materials to ensure that they explain to research participants the risks and benefits of submitting genomic data to NIH so that it can be shared with other investigators for secondary research use. NIH has developed a set of points for IRBs to consider when reviewing such genomic research proposals.[29]

Investigators seeking to download controlled-access data from an NIH data repository must sign a Data Use Certification Agreement[30] and abide by the NIH Genomic Data User Code of Conduct.[31] The Data Use Certification includes a series of data privacy and security requirements to which an investigator and his or her institution must agree.

This section of the report discusses the privacy and security safeguards incorporated in the GDS Policy. First, it provides some background on the Common Rule as well as the HIPAA Privacy and Security Rules, with which the Common Rule intersects. Both sets of standards have been

[26] C. Stephen Redhead, Specialist in Health Policy, wrote this section.

[27] National Institutes of Health, *NIH Genomic Data Sharing Policy*, August 27, 2014, http://gds.nih.gov/PDF/NIH_GDS_Policy.pdf.

[28] The Common Rule is the informal name given to core federal regulations governing the protection of human subjects in research supported or conducted by the federal government. The regulations were first promulgated by HHS at 45 C.F.R. Part 46, Subpart A.

[29] National Institutes of Health, *NIH Points to Consider for IRBs and Institutions in their Review of Data Submission Plans for Institutional Certifications*, revised May 31, 2011, http://gds.nih.gov/pdf/PTC_for_IRBs_and_Institutions_revised5-31-11.pdf.

[30] National Institutes of Health, *Model Data Use Certification Agreement*, May 30, 2014, version, http://gds.nih.gov/pdf/Model_DUC.pdf.

[31] National Institutes of Health, *Genomic Data User Code of Conduct*, revised April 2, 2010, http://gds.nih.gov/pdf/Genomic_Data_User_Code_of_Conduct.pdf.

criticized for their treatment of research data. Some privacy advocates complain that neither the Common Rule nor HIPAA adequately protects patient privacy, while researchers claim that HIPAA impedes their access to data and places limitations on its secondary use. There are also concerns about inconsistencies between the two sets of standards. Some of these concerns were addressed by HHS in a 2013 final rule that made numerous other amendments to the Privacy Rule pursuant to the Health Information Technology for Economic and Clinical Health (HITECH) Act.[32]

Common Rule

Under the Common Rule—the core federal regulations governing the protection of human subjects in government-supported research—research protocols must be approved by an Institutional Review Board (IRB) to ensure that the rights and welfare of the research subjects are protected.[33] The rule lists several criteria for IRB approval, including the requirement that researchers obtain the informed consent of their research subjects.[34] In addition, it sets out the types of information that must be provided to prospective research subjects during the informed consent process, including an explanation of the purpose of the research, a description of the research procedures, and a description of the risks and benefits of the research.[35]

An IRB may decide to waive the informed consent requirement if it determines that (1) the research poses no more than minimal risk to the subjects, (2) the waiver will not adversely affect the rights and welfare of the subjects; and (3) the research is not practicable without a waiver.[36]

While all forms of human subject research potentially involve privacy issues, the focus of the Common Rule (and IRB review) traditionally has been to protect the *safety* of individuals enrolled in clinical and other interventional research. But with the enormous growth in health data analytics, which often entails the secondary analysis of large databases of clinical information, the principal risk to research subjects increasingly is not physical harm but a loss of privacy.

The Common Rule's definition of human subject research includes the collection of individually identifiable information about the research participants. Obtaining de-identified information, by itself, does not constitute human subject research, and such activity is not bound by the Common Rule's requirements. Individually identifiable information is defined as information for which the identity of the subject "is or may readily be ascertained."[37] There is no explicit standard for de-identified information, which by implication is information for which the subject's identity is not readily ascertained.

The Common Rule includes two brief provisions that address privacy. First, it specifies that IRBs may only approve research that is judged to have "adequate provisions to protect the privacy of subjects and to maintain the confidentiality of data."[38] Second, the informed consent process must

[32] Department of Health and Human Services, Office of the Secretary, "Modifications to the HIPAA Privacy, Security, Enforcement, and Breach Notification Rules Under the Health Information Technology for Economic and Clinical Health Act and the Genetic Information Nondiscrimination Act; Other Modifications to the HIPAA Rules; Final Rule," 78 *Federal Register* 5566, 5609-5613, January 25, 2013.

[33] 45 C.F.R. §46.109.

[34] 45 C.F.R. §46.111(a)(4).

[35] 45 C.F.R. §46.116(a).

[36] 45 C.F.R. §46.116(d).

[37] 45 C.F.R. §46.102(f).

[38] 45 C.F.R. §46.111(a)(7).

include "a statement describing the extent, if any, to which confidentiality of records identifying the subject will be maintained."[39] The Common Rule does not elaborate on these provisions by providing additional guidance or defining any terms.

One final point about the Common Rule needs emphasizing, which is that it permits consent for corollary and future research. This consent can occur when the primary research study is paired with other activities, such as the creation of a research database or repository where information and specimens obtained from a research participant are transferred and maintained for future research. In such instances, an IRB may approve an informed consent document that asks research participants to allow future research on their identifiable information or specimens, provided the future research uses are described in sufficient detail to allow an informed consent.

HIPAA Privacy and Security Rules

The HIPAA Privacy and Security Rules established a set of federal standards to help safeguard personal health information.[40] The HIPAA Rules apply to health plans, health care clearinghouses, and health care providers, which are collectively referred to as "covered entities." The HIPAA Rules also apply to the business associates of covered entities. These are organizations with whom covered entities share health information to help carry out their activities and functions.

The Privacy Rule covers "protected health information" (PHI) in any form or format that is created or received by a covered entity.[41] The Privacy Rule includes a de-identification standard. Health information is considered de-identified if 18 specified types of identifiers are removed, or if a qualified expert, using accepted statistical methods, determines that the reidentification risk is "very small."[42] De-identified information that meets this standard is not subject to the Privacy Rule. In the broadest terms, the Privacy Rule prohibits a covered entity from using or disclosing PHI except as expressly permitted or required by the rule.[43] For all uses or disclosures of PHI that are not otherwise permitted or required by the rule, covered entities must obtain the individual's written authorization.

The Privacy Rule requires covered entities to adopt reasonable administrative, technical, and physical safeguards to protect PHI from unauthorized access, use, or disclosure.[44] The accompanying Security Rule—applicable only to PHI in electronic form (ePHI)—establishes a series of security standards that are both technology-neutral and scalable, based on the size and complexity of the organization.[45] The administrative standards include security management, workforce, and training, as well as procedures for dealing with security incidents. The physical standards include facility access and security, and workstation use and security. And the technical standards for protecting digital information include access controls, individual and entity authentication, and encryption.

[39] 45 C.F.R. §46.116(a)(5).

[40] The HIPAA privacy and security standards are codified at 45 C.F.R. Part 164.

[41] PHI is defined as individually identifiable information "created or received" by a covered entity that "relates to the past, present, or future physical or mental health ... of an individual; the provision of health care to an individual; or the past, present, or future payment for the provision of health care to an individual." 45 C.F.R. §160.103.

[42] 45 C.F.R. §164.514(b). The 18 types of identifiers include names; geographical and email addresses; social security, medical record, health plan, and account numbers; photographic images; and biometric identifiers.

[43] 45 C.F.R. §164.502(A).

[44] 45 C.F.R. §164.530(c).

[45] 45 C.F.R. §164.306.

Each security standard is accompanied by one or more implementation specifications. Some implementation specifications are required; for example, to meet the security management standard, each organization must conduct an accurate and thorough risk analysis.[46] Other implementation specifications are "addressable."[47] Organizations must assess each addressable specification to determine whether it is a reasonable and appropriate safeguard before deciding whether to adopt it.

Under the Privacy Rule, PHI may not be used or disclosed for research without authorization, with three exceptions. First, IRB may waive the authorization requirement based on a determination that (1) the use or disclosure of PHI involves no more than minimal risk to the privacy of the individuals; (2) the research could not practicably be conducted without a waiver; and (3) the research could not practicably be conducted without access to, and use of, the health information.[48] These criteria for waiving authorization for the use or disclosure of PHI for research are similar to the Common Rule criteria for waiving informed consent to participate in research. Second, PHI may be reviewed when necessary to prepare a research protocol or for a similar purpose to prepare for research.[49] Third, PHI of persons who have died may be used or disclosed if necessary for research purposes.[50]

In 2009, the Institute of Medicine (IOM) released a report on the Privacy Rule's impact on research.[51] The IOM concluded that the rule does not adequately protect the privacy of health information and impedes the conduct of important new research. The report found considerable variation in how organizations that collect and use health data are interpreting and following the Rule. It discussed the challenges in reconciling the Privacy Rule with other federal regulations—primarily the Common Rule—that govern human subject research. The report also examined inconsistencies between the Privacy Rule and the Common Rule, neither of which applies uniformly to all health research.

For example, the Privacy Rule generally prohibited combining an authorization with any other legal permission to create a "compound" authorization, unless it was for the same study. Thus, a Privacy Rule authorization for a specific research study could be combined with Common Rule informed consent to participate in the research. But any separate research activity, such as collecting specimens or data for a central research database or repository, would require its own authorization. Unlike Common Rule informed consent, Privacy Rule authorizations also had to be study-specific; authorizations for future research were prohibited.

In 2013, HHS modified its interpretation of the Privacy Rule to address some of the inconsistencies with the Common Rule.[52] Under the new interpretation, the Privacy Rule now

[46] 45 C.F.R. §164.308(a)(1)(ii)(A). The purpose of a risk analysis is to identify all the potential risks and vulnerabilities to the confidentiality, integrity, and availability of ePHI maintained by covered entities and their business associates. A risk analysis is the first and most important action a covered entity must take to comply with the HIPAA Security Rule. The results of the analysis should then guide all subsequent compliance actions.

[47] 45 C.F.R. §164.306(d)(3).

[48] 45 C.F.R. §164.512(i)(1)(i)-(ii).

[49] 45 C.F.R. §164.512(i)(1)(ii).

[50] 45 C.F.R. §164.512(i)(1)(iii).

[51] Institute of Medicine, *Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health Through Research*, Washington, DC, February 2009, http://www.nap.edu/openbook.php?record_id=12458.

[52] Department of Health and Human Services, "Modifications to the HIPAA Privacy, Security, Enforcement, and Breach Notification Rules Under the Health Information Technology for Economic and Clinical Health Act and the Genetic Information Nondiscrimination Act; Other Modifications to the HIPAA Rules; Final Rule," 78 *Federal Register* 5566, 5609-5613, January 25, 2013.

permits compound authorizations for any type of research activity (with limited exceptions) and allows authorizations for future research, provided the description of the future research uses is sufficiently clear that it would be "reasonable for an individual to expect that his or her protected health information could be used or disclosed for such future research."

Responsibilities of Researchers Submitting Genomic Data

NIH has developed a set of points for IRBs to consider when reviewing genomic research proposals that involve the submission of data to NIH.[53] The purpose of this document is to help inform and guide IRBs as they seek to determine, as required under the Common Rule, whether adequate data privacy protections are in place, and whether the informed consent process describes how data confidentiality will be maintained.

The NIH points-to-consider document includes background information on the GDS Policy and discusses both the benefits and risks of sharing genomic data through an NIH data repository. The document discusses several potential risks associated with the submission of genomic data to NIH and its subsequent release for secondary research. Those risks include the risk of identifying research participants, the risk of inadvertent or inappropriate use or disclosure of identifiable information, and the risk of disclosure in response to a request under the Freedom of Information Act (FOIA).

To reduce the risk of identification, investigators submitting genomic data to NIH-designated repositories are required to de-identify the data according to both the Common Rule and the Privacy Rule standards. As discussed earlier, only the Privacy Rule provides an explicit standard for de-identifying data. The submitting investigator should assign random, unique codes to the de-identified data and retain the identification keys.

While the NIH genomic data repository does not include individual identifiers (e.g., name, address, birth date, social security number), the agency recognizes that "technologies available within the public domain today, and technological advances expected over the next few years, make the identification of specific individuals from raw genotype-phenotype data feasible and increasingly straightforward."[54]

NIH encourages IRBs to consider whether an investigator has obtained a Certificate of Confidentiality from NIH as an additional layer of protection. A Certificate of Confidentiality protects investigators from being compelled to disclose information that would identify research subjects in any civil, criminal, administrative, legislative, or other proceeding.[55] This requirement can help promote participation in the research by adding an additional layer of privacy protection.

It is quite possible that the genomic research participants will be given a compound authorization that includes the informed consent materials (pursuant to the Common Rule) as well as a HIPAA authorization (pursuant to the Privacy Rule)—unless waived by an IRB—to allow an investigator to access medical information about the participants from their physicians and other health care providers. The HIPAA authorization form must include a description of the potential future uses of the data.

---

[53] National Institutes of Health, *NIH Points to Consider for IRBs and Institutions in their Review of Data Submission Plans for Institutional Certifications*, http://gds.nih.gov/pdf/PTC_for_IRBs_and_Institutions.pdf.

[54] Ibid., p. 5.

[55] Certificates of Confidentiality are issued pursuant to §301(d) of the Public Health Service Act (42 U.S.C. §241(d)). See http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/GetPdf.cgi?document_name=ConfidentialityCertificate.pdf.

Notwithstanding the Privacy Rule's requirements governing researchers' access to medical information about their research subjects, it is important to keep in mind that the research investigators themselves (and their institutions) are unlikely to be HIPAA-covered entities (i.e., health plans or health care providers). If researchers do not meet the definition of a covered entity, then they are not subject to the HIPAA privacy and security standards.

Responsibilities of Investigators Accessing and Using Genomic Data

Investigators and institutions seeking access to data from an NIH genomic data repository must submit a Data Access Request[56] along with a Data Use Certification.[57] The Data Use Certification specifies the terms and conditions for the research use of the data. For example, investigators must (1) follow all applicable federal, state, and local laws and regulations for handling genomic data, including IRB approval if required; (2) use the data only for the approved research; (3) not attempt to identify or contact the individual participants from whom the data were obtained; and (4) not share the data with anyone other than those listed in the Data Access Request.

The Data Use Certification also requires investigators (and their institutions) to agree to handle the data according to NIH's current dbGaP (database of genotypes and phenotypes) Security Best Practices.[58] These include, but are not limited to, the following IT security requirements: use of firewalls and updated anti-virus/anti-spyware software; use of security auditing/intrusion detection software; strong password policies; and encrypting data on portable devices. In general, investigators are required to keep the data secure and confidential and adhere to data management practices so that only authorized individuals gain access to the data.

Finally, investigators must notify NIH of any unauthorized data sharing, breaches of data security, or inadvertent data releases that may compromise data confidentiality within 24 hours of when the incident was identified.

As already noted, investigators and the academic and other institutions to which they belong are unlikely to be covered entities, in which case they are not bound by the HIPAA privacy and security standards. However, the dbGaP Security Best practices broadly overlap with the HIPAA technical security standards for protecting digital information and controlling access to it.

The Freedom of Information Act (FOIA)[59]

In 1966, Congress enacted the Freedom of Information Act (FOIA), which provides the public presumed access to executive branch information. FOIA established, for any person—corporate or individual, citizen or otherwise—presumptive access to existing, unpublished agency records on any topic.[60] In a "points to consider" memorandum regarding data sharing concerns, the NIH

[56] https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?page=login.

[57] National Institutes of Health, *Model Data Use Certification Agreement*, May 30, 2014, version, http://gds.nih.gov/pdf/Model_DUC.pdf.

[58] National Institutes of Health, *dbGaP Best Practice Requirements: Security Best Practices*, updated December 2, 2013, http://www.ncbi.nlm.nih.gov/projects/gap/pdf/dbgap_2b_security_procedures.pdf.

[59] This section was written by Wendy Ginsberg, Analyst in American National Government and Daniel J. Richardson, Research Assistant. FOIA is at 5 U.S.C. §552.

[60] For more information on FOIA, see CRS Report R41933, *The Freedom of Information Act (FOIA): Background, Legislation, and Policy Issues*, by Wendy Ginsberg.

stated that "datasets submitted to NIH" will be "U.S. Government records that are subject to" FOIA.[61] FOIA, however, specifies nine categories of information that may be exempted from the rule of disclosure, including trade secrets and information related to national security. Disputes between requesters and agencies over the accessibility of requested records may be settled in federal court or may be mediated in the Office of Government Information Services (OGIS).

The sharing of genetic and genomic data among private individuals, researchers, and the federal government has, at times, prompted concerns that the information, if collected or retained by a federal executive branch agency, could be subject to public release pursuant to FOIA. As noted above, although the information submitted to NIH is considered de-identified, researchers have demonstrated an ability to identify individual genomic or genetic material despite attempts to anonymize the data. Public release of the de-identified data, therefore, may generate unease about personal privacy protection as well as lead to calls for legislation further clarifying public access to such information.

Among FOIA's nine exemptions that permit agencies to withhold applicable records, two categories are more likely to affect the disclosure of genetic or genomic material:

- Exemption 3: data specifically exempted from disclosure by a statute other than FOIA if that statute meets criteria laid out in FOIA; and
- Exemption 6: Personnel, medical, or similar files, the disclosure of which would constitute an unwarranted invasion of personal privacy.

An example Exemption 3 statute, in the context of potentially withholding genetic material, might include 42 U.S.C. §242m(d) (a provision of the Public Health Service Act, as amended), which protects from public release certain information that would allow an individual to be identified if that information was collected for epidemiological or statistical activities. These types of Exemption 3 statutes are often referred to as b(3) exemptions because they are authorized in 5 U.S.C. §552(b)(3).

A second potentially applicable b(3) exemption is provided in 15 U.S.C. §3710a(c)(7)(a) (a provision of the National Defense Authorization Act for Fiscal Years 1990 and 1991), which protects from public release "trade secrets or commercial or financial information that is privileged or confidential ... obtained in the conduct of research ..."[62] Also, Exemption 3 would prohibit the disclosure of any information that is covered by future statutes passed by Congress. As a result, legislation enacted at any time that would specifically prohibit the release of genetic and genomic data may qualify as a b(3) exemption and could be used to withhold qualifying genetic material.

Exemption 6 of FOIA provides for the withholding of information that relates to personally identifiable information in "personnel and medical files, and similar files." The intention of this exemption is to allow agencies to withhold records that contain personally identifiable information, provided that the individual's interest in privacy outweighs the public interest in the record's release. According to the NIH "points to consider" memorandum, NIH "believes that the release of unredacted GWAS datasets ... would constitute an unreasonable invasion of personal

[61] National Institutes of Health, "Genome-Wide Association Studies (GWAS): NIH Points to Consider," http://gds.nih.gov/pdf/PTC_for_IRBs_and_Institutions_revised5-31-11.pdf.

[62] It is unclear, however, whether and in what contexts genetic or genomic material would legally qualify as a trade secret or commercial or financial information. For more information on the definition of *trade secret* in a legal context see CRS Report R43714, *Protection of Trade Secrets: Overview of Current Law and Legislation*, by Brian T. Yeh.

privacy under FOIA Exemption 6."[63] Moreover, Exemption 6 has been read historically as applying to information that can be linked to a particular individual.[64]

As discussed above, the ability to identify an individual using genetic research material that was previously not thought to be identifiable has been demonstrated. Where this data could be traced to specific individuals, it is possible that the potential release of the information would trigger a privacy interest of a degree that might warrant withholding under Exemption 6. The federal government, however, has maintained that the data collected are currently not identifiable, possibly removing any ability for an agency to apply an exemption that relies on the personal identification of an individual. Moreover, a determination that the data allows for an individual's identification would not necessarily permit the withholding of information; pursuant to FOIA, the privacy interests of the individuals who may be affected by the information's release would still need to be weighed against the public's interest in the information's disclosure.

In addition to these two exemptions, the specific facts surrounding each information request could trigger a number of other FOIA exemptions. For instance, genetic and genomic records could potentially relate to ongoing law enforcement activities (Exemption 7), inter-agency and intra-agency memorandums (Exemption 5), or confidential commercial and financial information (Exemption 4).

The Genetic Information Nondiscrimination Act (GINA, P.L. 110-233)[65]

In terms of the reidentification of research subjects, the Genetic Information Nondiscrimination Act of 2008 (GINA, P.L. 110-233)[66] would protect against discrimination based on information discovered about a research subject subsequent to his or her reidentification (e.g., genetic test results or family history). As described previously, one way to consider the security of data from a policy perspective is to consider it in terms of both controls on access to, and designation of appropriate uses of, the data. With respect to genetic information, GINA establishes prohibitions that affect primarily the appropriate use of genetic information, but that also address—to a lesser extent—access to genetic information. Specifically, GINA prohibits discrimination based on genetic information by both health insurers

How Does GINA Define Genetic Information?

GINA defines "genetic information" as follows: "The term 'genetic information' means, with respect to any individual, information about—(i) such individual's genetic tests, (ii) the genetic tests of family members of such individual, and (iii) the manifestation of a disease or disorder in family members of such individual."

The statute goes on to clarify that "[s]uch term includes, with respect to any individual, any request for, or receipt of, genetic services, or participation in clinical research which includes genetic services, by such individual or any family member of such individual."

The term "genetic information" excludes information about the sex or age of any individual.

Source: See 29 U.S.C. §1191b(d).

[63] The memorandum notes, however, that "FOIA affords requesters an opportunity to contest an agency's determination." See National Institutes of Health, "Genome-Wide Association Studies (GWAS): NIH Points to Consider," p. 6, http://gds.nih.gov/pdf/PTC_for_IRBs_and_Institutions_revised5-31-11.pdf.

[64] U.S. Department of Justice, Office of Information Policy, *Guide to the Freedom of Information Act*, "Exemption 6," pp. 4-5, http://www.justice.gov/sites/default/files/oip/legacy/2014/07/23/exemption6.pdf.

[65] Amanda K. Sarata, Specialist in Health Policy, wrote this section.

[66] For more information about GINA, see CRS Report RL34584, *The Genetic Information Nondiscrimination Act of 2008 (GINA)*, by Amanda K. Sarata and Jody Feder.

and employers. The reach of GINA's prohibitions is in part governed by its definition of the term "genetic information" (see "**How Does GINA Define Genetic Information?**" text box, above). Genomic sequence data would not necessarily be protected under GINA; instead, it would be the information uncovered secondary to analysis of the sequence data (i.e., a specific genetic test result) that would be protected under this statute.

GINA is divided into two main parts: Title I, which prohibits discrimination based on genetic information by health insurers, and Title II, which prohibits discrimination in employment based on genetic information. Title I of GINA amends the Employee Retirement Income Security Act of 1974 (ERISA), the Public Health Service Act (PHSA), and the Internal Revenue Code (IRC), through the Health Insurance Portability and Accountability Act of 1996 (HIPAA), as well as the Social Security Act (SSA), to prohibit group health plans and health insurance issuers from engaging in genetic discrimination. Broadly, GINA prohibits group health plans and health insurance issuers from engaging in three practices: (1) using genetic information about an individual to adjust a group plan's premiums, or, in the case of individual plans, to deny coverage, adjust premiums, or impose a preexisting condition exclusion; (2) requesting, requiring, or purchasing genetic information for underwriting purposes or prior to enrollment; and (3) requiring or requesting genetic testing.

While the first prohibition addresses the *use* of the information, the second addresses both *access* to and *use* of the information, and the last addresses *access* to the information.

The health reform law (ACA, P.L. 111-148, as amended) contains provisions that may overlap to some extent with those in Title I of GINA. In evaluating the interaction of these two statutes, one may argue that it is possible to read these statutes together as establishing non-conflicting limitations on insurance premiums. Although GINA prohibits using genetic information to determine health coverage and insurance premiums for individuals or groups, the ACA specifically defines the factors on which insurers may predicate issuance of coverage or determination of premiums. The relevant provisions in GINA and the ACA are not identical in scope; however, the provisions of the ACA may obviate some of the requirements of GINA. Importantly, in terms of access to genetic information, there does not seem to be a comparable provision in the ACA to GINA's prohibition on group health plans and health insurers from requiring an individual or family member to undergo a genetic test.[67]

Title II of GINA includes provisions that address both access to and appropriate use of genetic information by employers. Specifically, Title II of GINA prohibits discrimination in employment because of genetic information and, with certain exceptions, prohibits an employer from requesting, requiring, or purchasing genetic information. The law prohibits the use of genetic information in employment decisions—including hiring, firing, job assignments, and promotions—by employers, unions, employment agencies, and labor-management training programs. Title II outlines exceptions whereby an employer may lawfully acquire genetic information (e.g., through inadvertent requests, wellness programs, or DNA analysis for law enforcement purposes, among others). However, even if genetic information is acquired through these exceptions, the employer may not use it to discriminate.

If genetic information, as defined by GINA, becomes more easily accessible because of the ability to link genomic sequence data with specific individuals, then policymakers may consider expanding GINA's applicability to broaden its protections regarding the use of genetic information. In other words, if access to the data is becoming more difficult to control due to

[67] For more information on the interaction of the ACA and GINA, see CRS Report R41314, *The Genetic Information Nondiscrimination Act of 2008 and the Patient Protection and Affordable Care Act of 2010: Overview and Legal Analysis of Potential Interactions*, by Amanda K. Sarata and Jennifer A. Staman.

advances in technology or the favoring of countervailing policy goals (e.g., promoting advances in research), then an alternative policy approach is to strengthen requirements governing how the data are able to be used lawfully. This may be done by adjusting the requirements themselves to make them stricter, by broadening the applicability of the existing requirements to additional settings or arrangements, or by a combination of both of these approaches. GINA, for example, does not apply to life, disability, or long-term care insurance, nor does it apply to TRICARE, the Indian Health System (IHS), the Veterans Health Administration, or the Federal Employees Health Benefits Program (FEHB).[68]

Author Information

Amanda K. Sarata, Coordinator
Specialist in Health Policy

C. Stephen Redhead
Specialist in Health Policy

Wendy Ginsberg
Analyst in American National Government

Daniel J. Richardson
Research Assistant

Disclaimer

[68] R. C. Green, D. Lautenbach, A. L. McGuire, "GINA, Genetic Discrimination, and Genomic Medicine," *New England Journal of Medicine*, vol. 372, no. 5, January 29, 2015.